11

1

THE EFFECT OF KNOWLEDGE ON THE CALIBRATION OF PROBABILITY ASSESSMENTS

OREGON RESEARCH INSTITUTE

Sarah Lichtenstein
Baruch Fischhoff

ADVANCED
DECISION TECHNOLOGY
PROGRAM

CYBERNETICS TECHNOLOGY OFFICE
DEFENSE ADVANCED RESEARCH PROJECTS AGENCY
Office of Naval Research • Engineering Psychology Programs

The objective of the Advanced Decision Technology Program is to develop and transfer to users in the Department of Defense advanced management technologies for decision making. These technologies are based upon research in the areas of decision analysis, the behavioral sciences and interactive computer graphics. The program is sponsored by the Cybernetics Technology Office of the Defense Advanced Research Projects Agency and technical progress is monitored by the Office of Naval Research – Engineering Psychology Programs. Participants in the program are:

Decisions and Designs, Incorporated
The Oregon Research Institute
Perceptronics, Incorporated
Stanford University
The University of Southern California

Inquiries and comments with regard to this report should be addressed to:

Dr. Martin A. Tolcott
Director, Engineering Psychology Programs
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

or

LT COL Roy M. Gulick, USMC
Cybernetics Technology Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

***The views and conclusions contained in this document are those of the author(s) and should not be interpreted as necessarily* representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government. This document has been approved for public release with unlimited distribution.**

TECHNICAL REPORT DDI-4

THE EFFECT OF KNOWLEDGE ON THE CALIBRATION OF PROBABILITY ASSESSMENTS

by

Sarah Lichtenstein and Baruch Fischhoff

Sponsored by

Defense Advanced Research Projects Agency
Contract N00014-76-C-0074
ARPA Order No. 3052
Under Subcontract from
Decisions and Designs, Incorporated

August, 1976

OREGON RESEARCH INSTITUTE
P.O. Box 3196
Eugene, Oregon 97403
(503) 484-2123

SUMMARY

Introduction

Many, if not most, probability assessments come in the form of statements like, "I am XX% certain that the answer to this question is Y." A series of 5 experiments exploring the validity of such probability judgments are described in this report. Such judgments appear to have a moderate but systematic bias which is surprisingly insensitive to some factors (like the intelligence or expertise of the assessor) and surprisingly sensitive to others (the difficulty of the question). The implications of these results for decision making are discussed.

Background and Approach

Most important decisions involve uncertainty. That uncertainty is typically quantified in subjective probability assessments of the form "I am XX% certain that proposition Y is true." Proposition Y might be: "There will be no major outbreaks on Cyprus before the end of the year" or "Appropriation Z will be approved as requested." Although it is generally impossible to assess the validity of any one such probability assessment, a set of such estimates can be evaluated according to their "degree of calibration." They will be perfectly calibrated if the XX% of the propositions assigned probability XX turn out to be true (e.g., 50% of those given a .50 chance of being true). Any systematic bias in such probability estimates could lead to inaccuracies in decisions relying on them.

Five experiments, involving over five hundred people, studied the calibration of subjective probability assessments assigned to propositions regarding a wide variety of general knowledge questions. For each question, people chose one of two possible answers as the correct one, and then gave the probability that their answer was correct.

Findings

1. People's probability estimates show moderate validity for all but the most difficult tasks.

2. The most common sources of invalidity are: (a) overconfidence: people believe that they know more than they actually do; (b) insensitivity: people believe that they can discern finer distinctions in their own subjective uncertainty than they actually can.

3. People are no better calibrated when dealing with questions in their own area of expertise than when dealing with general knowledge questions.

4. Intelligence of the assessor has no effect on calibration.

5. Calibration changes markedly with questions of different difficulty. Although people are typically overconfident, that overconfidence increases as questions get more difficult and changes to underconfidence with the easiest questions.

Several simulations were conducted to make certain that these conclusions were not artifactual.

Implications

Sophisticated decision analyses typically include sensitivity analyses which show how sensitive their conclusions are to errors in the probability and utility estimates on which they are based. Results of the present experiments show what range of errors should be included in sensitivity analyses. Further work is needed to see if different kinds of questions and different ways of asking for probabilities produce similar errors--and if there is one best way to elicit probabilities.

Any systematic bias in probability assessment suggests the following intriguing possibility: instead of using the biased probabilities that people give us, why not use corrected estimates that take known biases into consideration. If, for example (as shown in "The Certainty Illusion" by Slovic, Fischhoff, and Lichtenstein), people should be saying .90 when they say .99, why not treat any estimate of .99 as though it were actually .90. The exact correction would presumably vary from situation to situation. Findings (3) and (4) make this situational adjustment easier by showing that intelligence and expertise are two factors that need not be considered. Finding (5) poses a real problem for this approach: If the error in calibration depends on the difficulty of the questions, then efficient correction requires knowledge of question difficulty. To know that, we must know the right answers to the questions. If we know that (e.g., if we know that there will be an outbreak on Cyprus before the end of the year), then we have no need for probabilities. The report suggests one way of capitalizing on changes in the probabilities people use to provide an indicator of how difficult the questions are--and what sort of correction factor should be used. However, it concludes that the best way to resolve this problem is to improve probability estimation by training, and thus do away with the correction problem.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENT

Support for this research performed by Oregon Research Institute, was provided by the Advanced Research Projects Agency of the Department of Defense and was monitored under Contract N00014-76-C-0074 with the Office of Naval Research, under subcontract from Decisions and Designs, Inc.

THE EFFECT OF KNOWLEDGE ON THE CALIBRATION OF PROBABILITY ASSESSMENTS

INTRODUCTION

Dealing with uncertainty is a central challenge in our day-to-day lives. In order to manage our affairs effectively, we must make predictions about the future behavior of individuals, groups, social systems, economies and international engagements. Reflecting this situation, subjective probabilities, the numerical expression of our predictions, have found their way into psychological theories of such diverse phenomena as motivation (Feather, 1959; Weiner, 1974), attitudes (Fishbein, 1967), personality attributions (Jones & Davis, 1965), decision making (Edwards & Tversky, 1967), choice behavior (Krantz, Luce, Suppes & Tversky, 1974), and gambling (Cohen, 1960). Subjective probabilities are also an integral part of sophisticated techniques like cost-benefit analysis and decision analysis that are used heavily in both business and social contexts (e.g., Atomic Energy Commission, 1975; Raiffa, 1968; Slovic, Kunreuther, & White, 1974).

The quality of people's probability assessments sets an upper limit on the quality of their functioning in uncertain environments. Knowing how good people are at assessing probabilities clearly has both theoretical and applied importance.

One approach to validating probability assessments is to restrict one's attention to situations in which a "correct" probability can be consensually defined, for example, situations where extensive frequentistic data are available and probabilities are essentially estimates of relative frequencies. Peterson and Beach (1967) reviewed a number of studies that adopted this approach and found that people can estimate relative frequencies quite well. More recently, however, Tversky and Kahneman (1973) have suggested systematic biases that may be present in such judgments.

For many tasks, however, a consensually defined "correct" answer is unavailable. This is particularly true for probabilities reflecting judges' degrees of belief in propositions concerning "unique" events (e.g., what is the probability that Portugal will withdraw from NATO within six months?) or the judge's knowledge about specific items of information (e.g., what is the probability that absinthe is a precious stone?). Such judgments reflect a degree of confidence entirely internal to the judge. Even if we know that Portugal did not withdraw from NATO during the period specified, or that absinthe is a liqueur, we can say nothing about how adequately the judge assessed and reported his or her own uncertainty. There is no way to evaluate an isolated judgment of this type.

Often, however, the judge makes many such responses, assessing the probability of many different unique events occurring or propositions being true. Over such a set of judgments, validity can be sought. One method of evaluating the quality of a set of probability judgments is to look at the internal consistency or coherence of the set. To be valid, subjective probability judgments must follow the axioms of the probability calculus. For example, since the two propositions given above are independent, the probability of both being true ("Portugal will withdraw from NATO" _and_ "absinthe is a precious stone") should be equal to the product of the probabilities of each being true. Wyer (1974) adopted this approach in a large number of studies and found a good deal of evidence of inconsistency, perhaps the most interesting aspect of which was a tendency to overestimate the likelihood of compound events. Internal consistency is a necessary condition for the validity of individual probability estimates, but it is not sufficient. Large systematic biases may exist in entirely consistent judgments.

A more direct method for evaluating the validity of a judge's assessments is to look at what we will call his or her degree of calibration. Assume that the true outcome of every proposition in the set is eventually known (by waiting six months to see what happens to Portugal, or by looking up absinthe in a dictionary). Then a judge is perfectly calibrated if, over the long run, for all propositions assigned the same probability, the proportion that are true is equal to the probability assigned. Thus, across that subset of answers to which the perfectly calibrated assessor assigns a probability of being correct of .7, 70% should be correct, and for all proportions to which .8 is assigned, 80% should be correct. For an assessor producing a large number of responses, one may group like responses and observe the hit rate for each subgroup. A graph showing the hit rate (percent correct) for each probability response is called a "calibration curve." Calibration curve A in Figure 1 reflects underconfidence: whenever such a person says .7, 88% of the answers are correct--such people know more than their responses indicate. Curve B, the diagonal, represents perfect calibration. Curve C represents overconfidence; for example, only 47% of all the events to which the judge responds .7 are indeed correct.

While a number of investigators have studied calibration, the only consistent finding has been that judges tend to be overconfident (for a review of this literature, see Lichtenstein, Fischhoff, & Phillips, 1976).

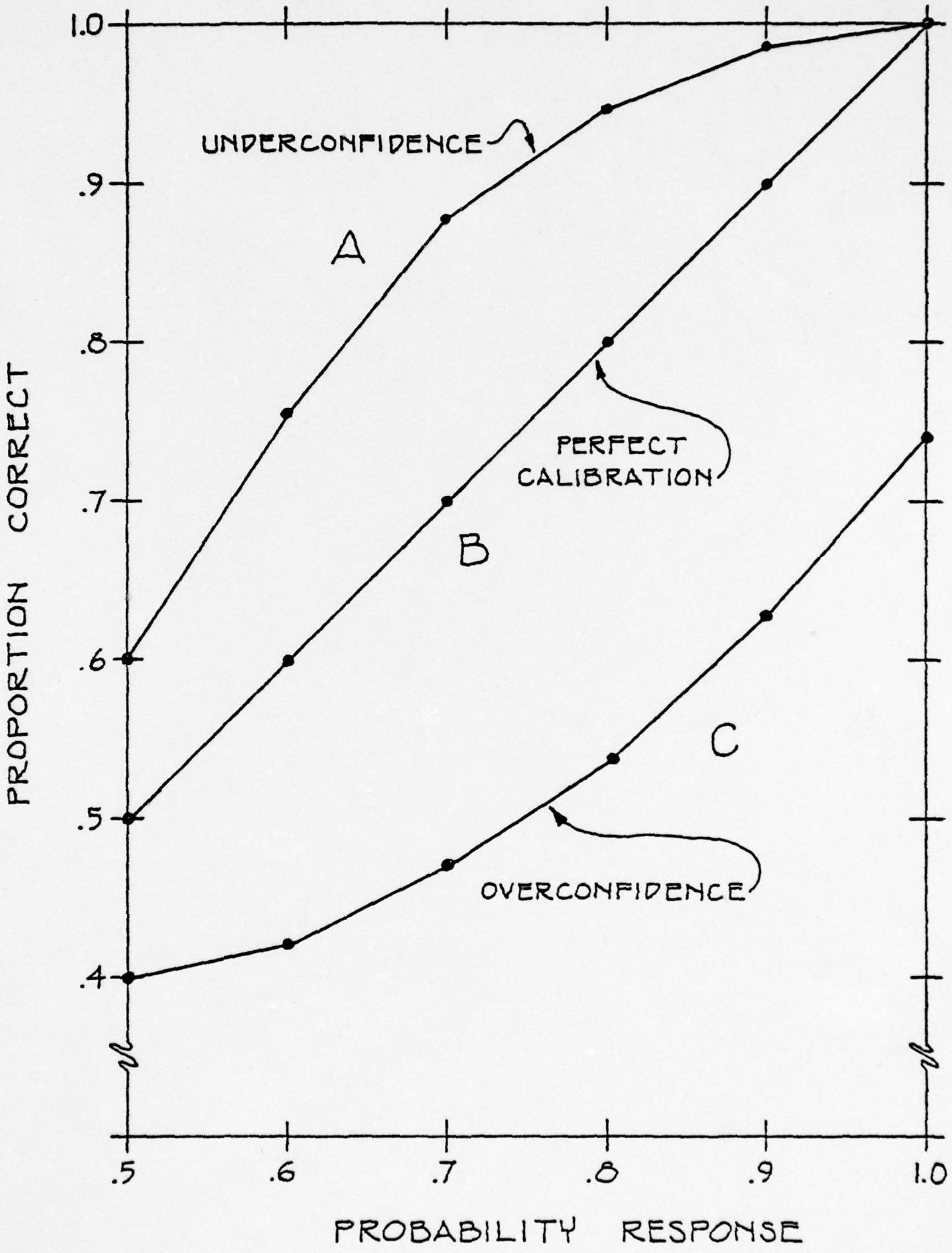


Figure 1

Exemplar calibration curves

The present studies constitute a systematic look at how well people are calibrated and what affects their degree of calibration. In particular, we want to know whether the amount of knowledge a judge possesses about the content of the propositions being assessed affects his or her calibration. Earlier studies (Adams & Adams, 1961; Clarke, 1960; Pitz, 1974; and Pollack & Decker, 1958) have reported some evidence that people who know more are better calibrated. The work reported here provides replication, clarification and extension of these findings.

All Experiments

Certain features shared by all experiments are reported here to avoid repetition.

Subjects. Except for Experiment 4, all subjects were paid volunteers who responded to advertisements in the University of Oregon student newspaper. Except for Experiment 4, the reported task was performed as part of a two-hour group session along with several other judgmental tasks. Group size varied from 25 to 48 persons.

Tasks. All test items were dichotomous items with the general form "Absinthe is (a) a precious stone, (b) a liqueur." In all experiments, subjects made two responses to each item. First, they chose one of the two alternatives as their best guess at the correct alternative. Second, they indicated with a number from .5 to 1.0 the probability that their choice was correct.

Measures. For each experiment, we report: (1) the percentage of questions for which the correct alternative was selected; (2) subjects' mean probability response; and (3) a calibration curve.

Calibration curves were constructed by grouping (over subjects and items) all the responses in the ranges .50-.59, .60-.69, .70-.79, .80-.89,

.90-.99, and 1.00. The mean response for each grouping is plotted against the percent correct (hit rate) associated with those responses.

No Knowledge

Experiments 1a and 1b investigated the calibration of subjects with severely limited knowledge.

Experiment 1a

Method. Each of 92 subjects was asked to decide, for each of 12 small drawings, whether the artist was a European child or an Asian child, and to estimate the probability that their selection was correct. Each set contained six drawings made by children from European countries and six drawings from Asian countries, all taken from Kellogg (1970), who had selected them to illustrate her thesis that children's drawings are the same the world over. This suggested that discrimination according to national origin would be very difficult. The test session was preceded by a brief study period in which the subjects were informed of Kellogg's thesis.

Results. As expected, the subjects had difficulty with this task. Only 53.2% of their 1104 answers were correct. Their probabilistic responses, however, indicated undue confidence, with a mean response of .677. The calibration curve shown in Figure 2 strongly suggests that these subjects were unaware of how little they knew. There is no relationship between their probability responses and the associated hit rates.

Experiment 1b

Method. Sixty-three subjects were taught how to read the stock market charts for individual companies provided by the weekly Standard and Poor report, Trendline. After the instruction period, they were given charts of twelve stocks with data for the period from July 9, 1974 to February 14, 1975. For each stock, they were asked to indicate whether its March 22 closing price was higher or lower than that of February 14. Each of four test sets included six stocks that had increased and six that had decreased over the period, chosen at random from all stocks that appeared in Trendline for February 14, 1975. Global market indices (e.g., Dow-Jones) were similar for February 14, the last day shown on the charts, and March 22, the target date, indicating that the market as a whole neither increased nor decreased during this period.

Results. Again, the task was too difficult for subjects to perform adequately. Only 47.2% of their choices were correct. Again, they overestimated their knowledge, providing a mean probability of .654. The calibration curve shown in Figure 2 shows the same insensitivity of probability judgments to level of knowledge found in Experiment 1a.

Comment. The lack of calibration evinced by the subjects in these two studies does not logically follow from their lack of knowledge. Subjects would have been quite well calibrated had they always given a probabilistic response of .5. This would have resulted in but one data point on the calibration curve for each experiment, but that point would have fallen reasonably close to the perfect calibration line. Only 7 of the 155 subjects in Experiments 1a and 1b acknowledged the limits of their own knowledge by following this strategy.

A Little Knowledge

Will a small amount of knowledge improve calibration? Experiment 2 was designed to investigate this possibility by partially training subjects to make the requisite discrimination.

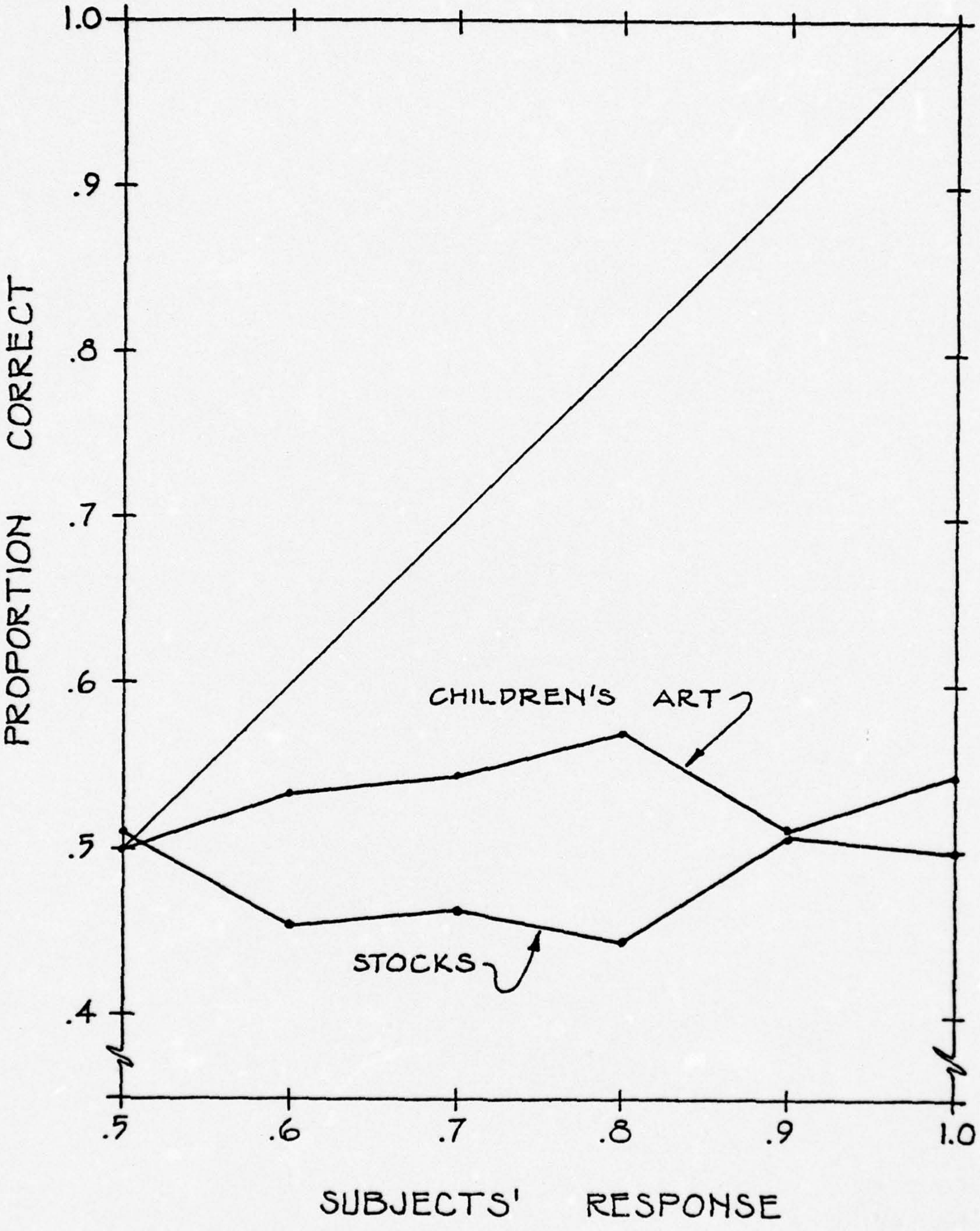


Figure 2

Experiment 1: Calibration with no knowledge

Experiment 2

Method. The stimuli were examples of the Latin phrase, "Mensa mea bona est," handwritten by either European or American adults. Twenty specimens were chosen on the basis of a pretest of 20 American subjects who were asked to sort 100 such specimens into two piles, American and European. The percent of correct identifications for the 20 specimens chosen for the experiment ranged from .40 to .60.[1] These 20 specimens were randomly divided into two sets of 10, each of which included 5 European and 5 American specimens. One set was used as training stimuli; the other was used as test stimuli. This random division was performed four times, producing four paired sets of training and test stimuli.

Two of four groups of subjects (N = 52) received training on this task. In the training phase, they were asked to study for five minutes the ten training stimuli, each correctly labeled. Immediately following this rudimentary training, the ten test stimuli were presented. For each, the subjects were asked to indicate whether the specimen was European or American, and to assess the probability that their answer was correct. They were not told how many of the ten test stimuli were American.

The procedure for the two groups of untrained subjects (N = 57) was identical except that the specimens they studied in the first phase were not labeled as to country of origin.

Results. Training was moderately successful; the trained subjects correctly identified 71.4% of the specimens, compared with 51.2% for untrained subjects. The mean responses were .779 for the trained group,

[1] We are grateful to Lewis Goldberg, out of whose files we stole, without his knowledge, the handwriting specimens and the pretest results.

.653 for the untrained group. As Figure 3 shows, trained subjects not only knew more, but also were better calibrated; the untrained subjects, as in Experiment 1, showed no evidence of calibration.

Different Levels of Knowledge

The suggestion that greater knowledge improves calibration was further explored in Experiment 3.

Experiment 3

Method. The stimuli were 150 general knowledge items with highly varied content (e.g., Aden was occuped in 1839 by the [a] British, [b] French; Bile pigments accumulate as a result of a condition known as [a] gangrene, [b] jaundice). One hundred twenty subjects each responded to 75 items drawn from a pool of 150 items; 25 of the items received 80 responses, 100 items received 60 responses, and 25 items received 40 responses.

Results. Figure 4 presents the calibration curve over all 9,000 responses. It is substantially flatter than it should be. The hit rates associated with the responses .50 and .60, and with .70 and .80, were virtually identical. Subjects generally overestimated the extent of their knowledge, getting 63.8% of the answers correct, but assigning a mean probability of .724.

The subjects were divided into three subgroups according to how knowledgeable they had been: the best subjects (40 subjects with 51 or more correct answers out of 75), the middle subjects (39 subjects with 46 to 50 correct answers), and the worst subjects (41 subjects with fewer than 46 correct answers). Separate analyses were performed for each group. Calibration curves appear in Figure 5, with the corresponding

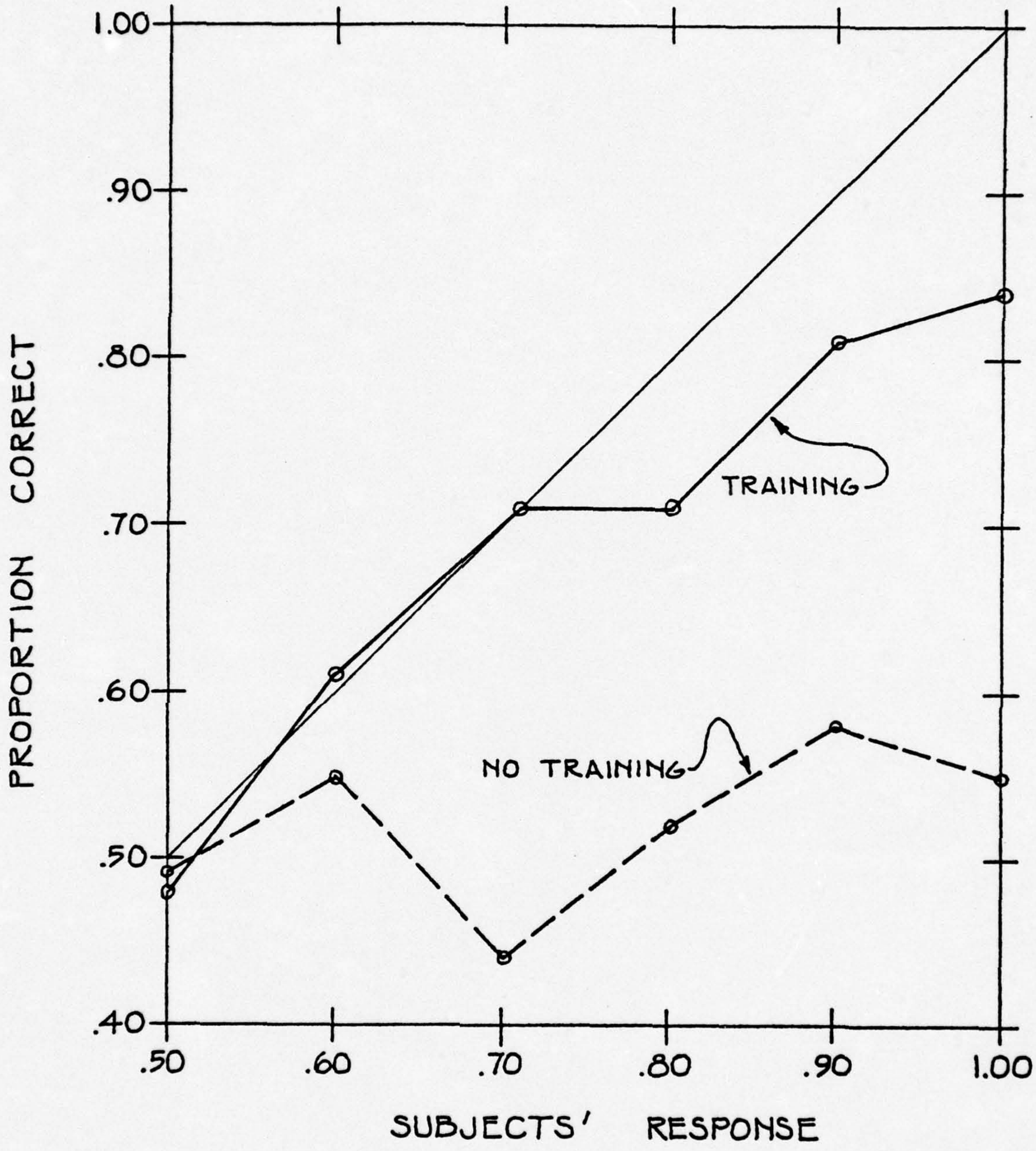


Figure 3

Experiment 2: "Mensa mea bona est." Effects of training on calibration

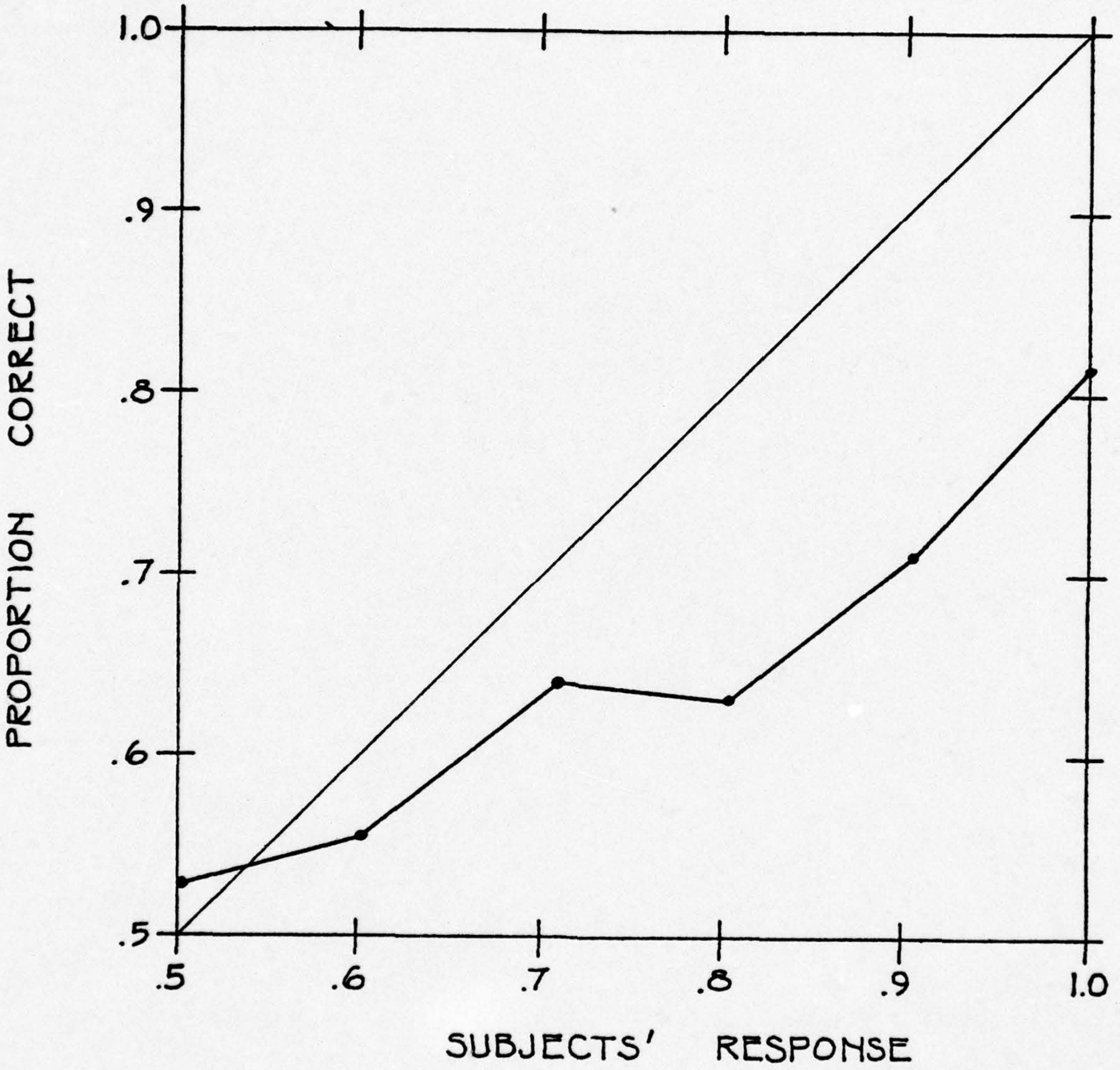


Figure 4

Experiment 3: Overall calibration curve.

General knowledge items, regular subjects

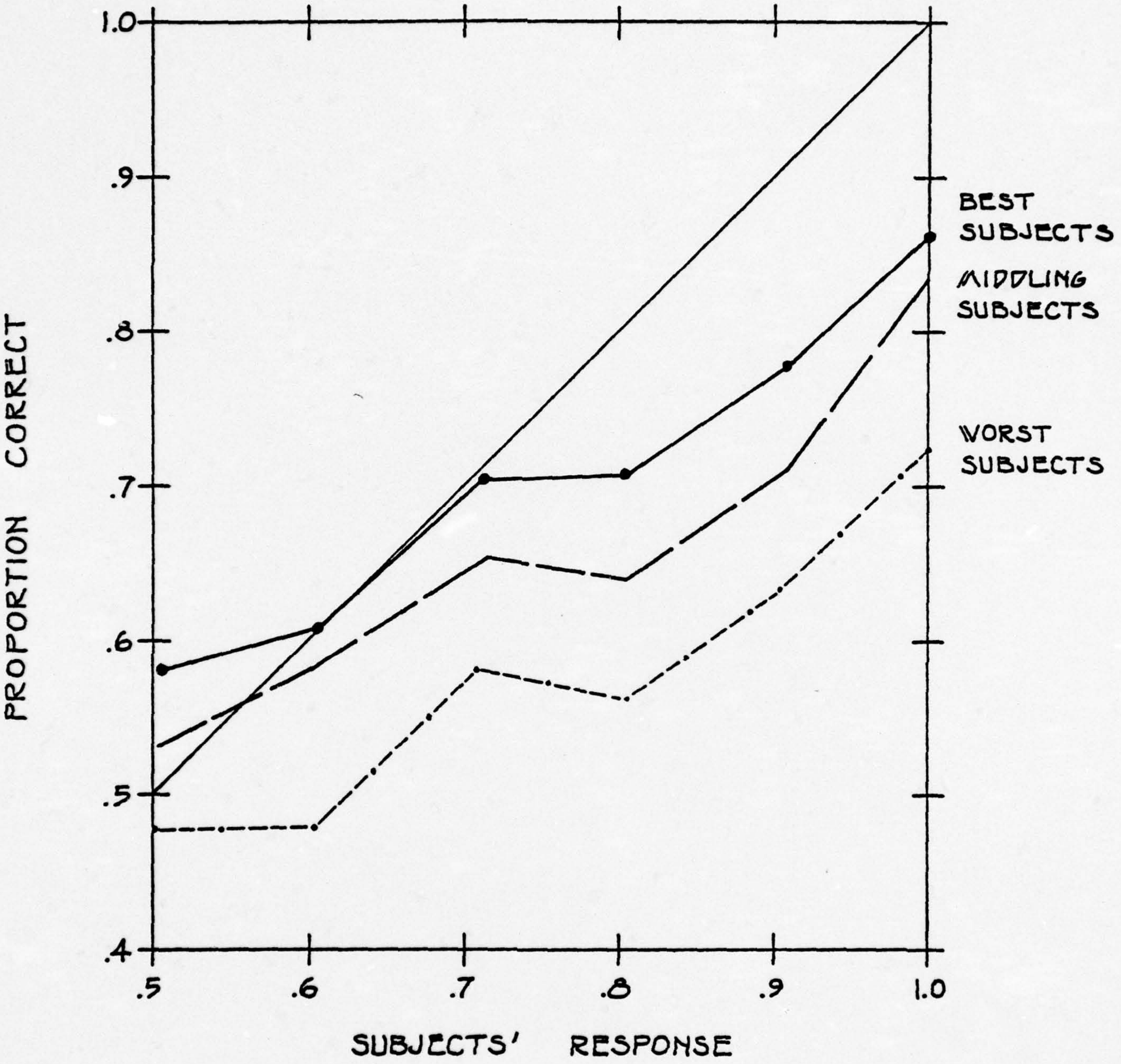


Figure 5

Experiment 3 again: Best vs. Worst subjects

TABLE 1

Summary Table of Calibration Statistics for Experiments 3 and 4

	Number of Responses	Percent Correct	Mean Response
Exp. 3			
All subjects	9000	.638	.724
By subject:			
Best 40 subjects	3000	.714	.743
Middle 39 subjects	2925	.643	.711
Worst 41 subjects	3075	.560	.706
By subject x item:			
Best subjects-easy items	1532	.847	.796
Middle subjects-easy items	1472	.800	.747
Worst subjects-easy items	1516	.695	.733
Best subjects-hard items	1468	.576	.716
Middle subjects-hard items	1453	.483	.674
Worst subjects-hard items	1559	.429	.679
Exp. 4			
All subjects	5000	.779	.784
By subject x item:			
Best subjects-easy items	1450	.923	.862
Worst subjects-easy items	1450	.847	.820
Best subjects-hard items	1050	.655	.705
Worst subjects-hard items	1050	.512	.681

statistics in Table 1. These data strongly suggest that the more one knows, the better one's calibration. All groups tended to overconfidence, but the most knowledgeable subjects showed the least overconfidence.

Dividing responses according to item difficulty rather than subject proficiency produced much the same result (not shown). The calibration curve for the easiest items was considerably closer to the identity diagonal than that for the most difficult questions.

Pushing this idea one step further, one might ask how well calibrated were the best subjects on the easiest items? Here, we might expect to find the best calibration. The data of Experiment 3 were re-analyzed to investigate this possibility. Items were sorted into two groups according to the percentage of subjects answering them correctly: easy items (67% or more correct) and hard items (less than 67% correct). Each of the three groups of subjects was calibrated for each of the two groups of items, to produce the six calibration curves shown in Figure 6. Summary statistics are shown in Table 1.

Despite some irregularities in these calibration curves due to the reduced number of responses per data point, a pattern of roughly parallel lines emerged. With low knowledge, substantial overconfidence occurred. However, when the percentage of correct answers was high (85% for the best subjects on the easy items and 80% for the middle subjects on the easy items), substantial underconfidence was seen (e.g., 75% of the .60 responses were correct). Calibration appears to change with increased knowledge, but not necessarily for the better.

Experiment 4

Although conducted for a somewhat different purpose (see below), Experiment 4 affords a replication of the above analysis.

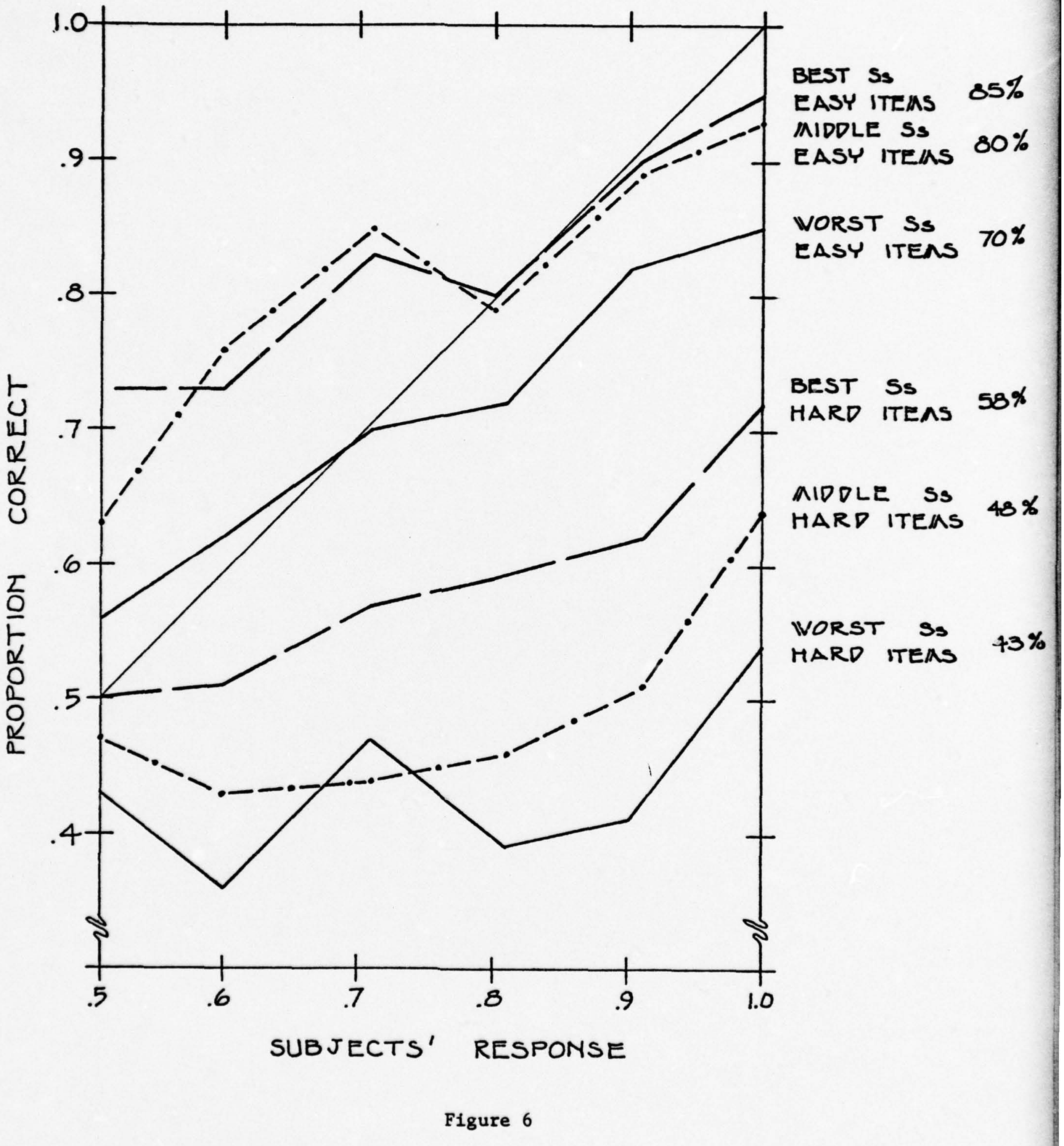


Figure 6

Experiment 3 yet again.

Calibration split six ways, by subjects and by items

Method. All on-campus graduate students in the Psychology Department of the University of Oregon were asked to participate in this experiment. Packets with stimuli and instructions were sent to all 64 graduate students; 50 were returned completed.

The stimuli were 50 general knowledge items (30 of those used in Experiment 3 and 20 additional, similar items) and 50 specially-written items dealing with psychology (e.g., the Ishihara test is [a] a perceptual test, [b] a social anxiety test; Anna Freud is Sigmund Freud's [a] oldest child, [b] youngest child). The two types of items were randomly intermixed in the stimulus package.

Results. Separate calibration curves are shown in Figure 7 for four subsets of responses obtained by splitting the subjects into best and worst at the median (74.5%) of the distribution of percentage correct, and splitting the items into easy (at least 75% correct; 58 items) and hard (fewer than 75% correct; 42 items). Summary statistics are given in Table 1. For these analyses, no distinction was made between general knowledge and psychology items. The same pattern of almost parallel lines found in Figure 6 emerged from these data.

Effects of Chance Fluctuations

The analytic technique used in Experiments 3 and 4, in which the data were divided into subsets as a function of item difficulty and subjects' performance, is vulnerable to random fluctuations which could artifactually produce separation between the calibration curves for the subsets. Assume that our subjects were equally knowledgeable and identically calibrated. In any sample of their responses, some will probably appear more knowledgeable by chance. The same chance factors that led them to have a higher

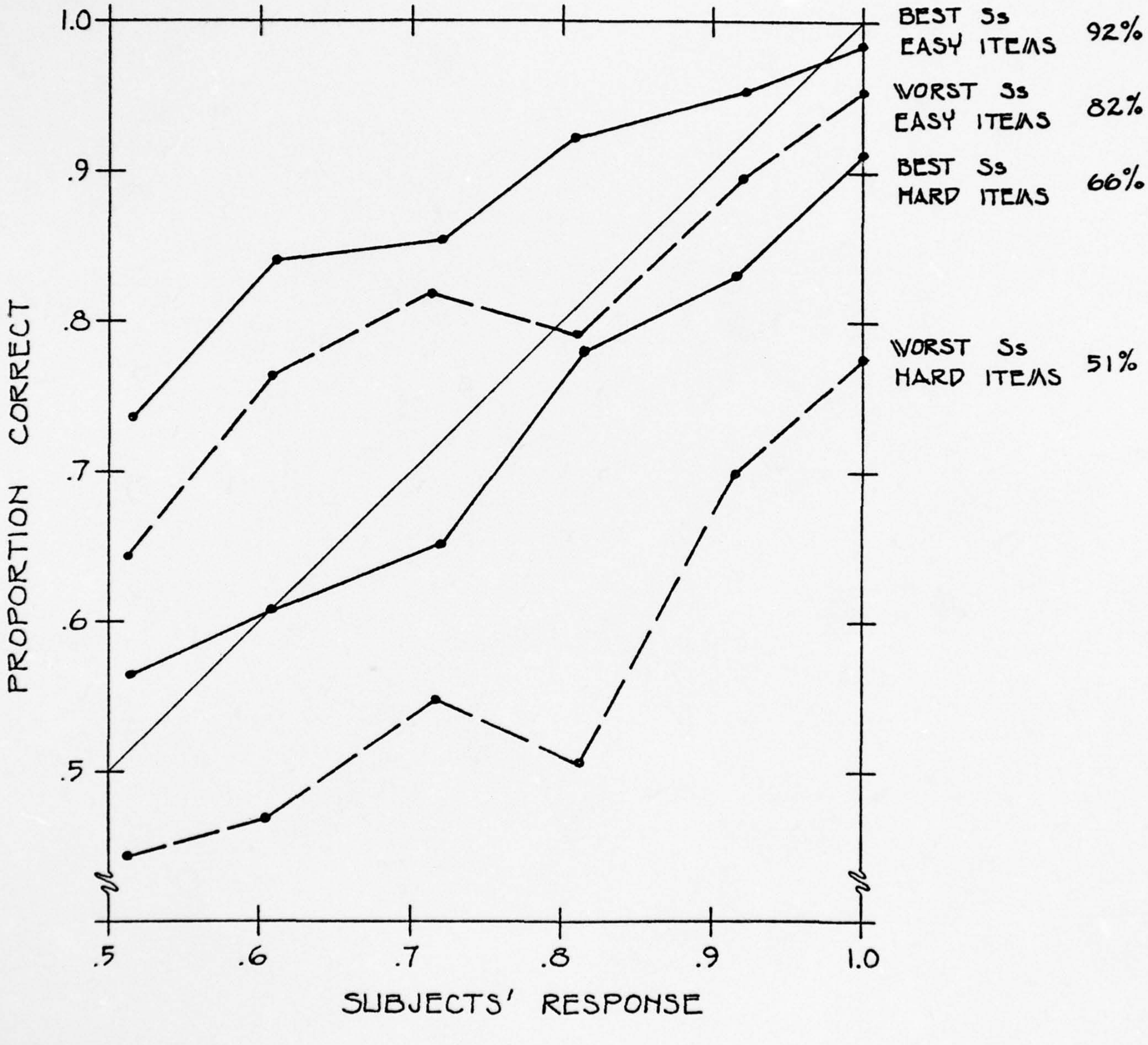


Figure 7

Experiment 4: Replication of the results of Figure 6, this time with graduate students in psychology overall percent correct will also lead them to have a higher hit rate for their responses of .5, .6, etc., and thus have an elevated calibration curve. The extent to which such chance factors could lead to differences in calibration was examined by simulating the results of Experiment 4. For the simulation, all subjects were assumed to have exactly the same calibration, which was taken as the actual calibration derived from pooling their responses to all 5,000 items (100 items for each of 50 subjects). Subjects' original probability responses were retained in the simulation. For each response, the correctness of the chosen alternative was simulated in accordance with the overall calibration curve. For example, since in the real data 86% of the .90 responses were correct, in the simulated data each response of .90 received a simulated outcome, either correct--with a probability of .86, or incorrect--with a probability of .14. These simulated data (the original probability responses with randomly chosen outcomes) were then partitioned into four subsets--best and worst subjects; easy and hard items. The calibration for each subset was computed. The entire simulation was repeated 50 times. Figure 8 shows the average calibration curves across the 50 replications. Figure 8 is directly comparable to Figure 7; it is based on the same data except for the assumption that all subjects have exactly the same calibration. The amount of separation between the calibration curves in Figure 8 is due solely to chance fluctuations. This separation is much smaller than the separation found in the original data (Figure 7). We reject the hypothesis that in Figure 7 all subjects on all items had the same calibration.

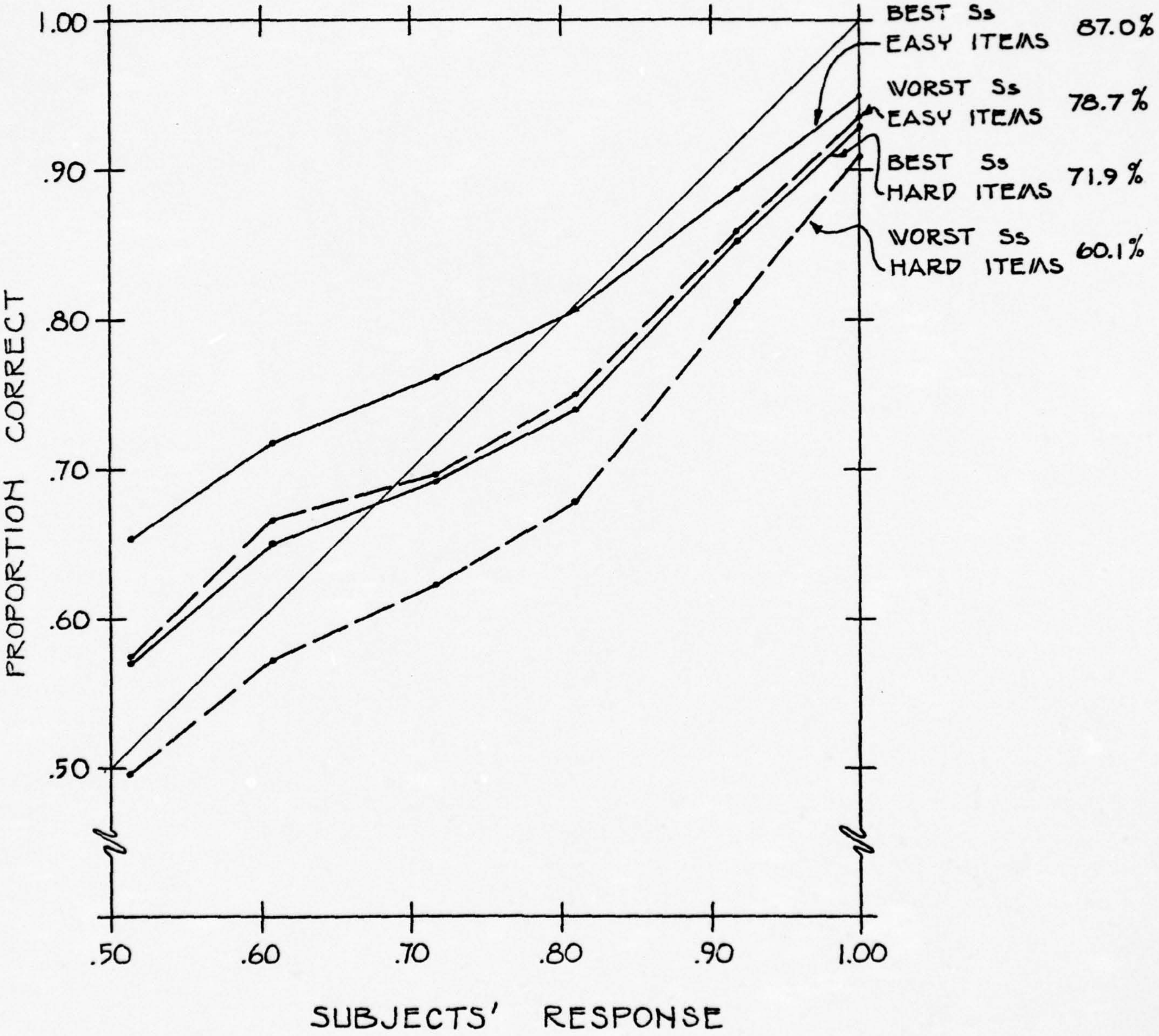


Figure 8

Results of a simulation to parallel the findings of the previous figure

Tests Varying in Difficulty versus Sub-tests Varying in Difficulty

The previous experiments analyzed subsets of items actually contained in a single test. It may be that some adaptation to the overall difficulty of the test might account for the observed overestimation with hard items and underestimation with easy items. This possibility was explored in Experiment 5.

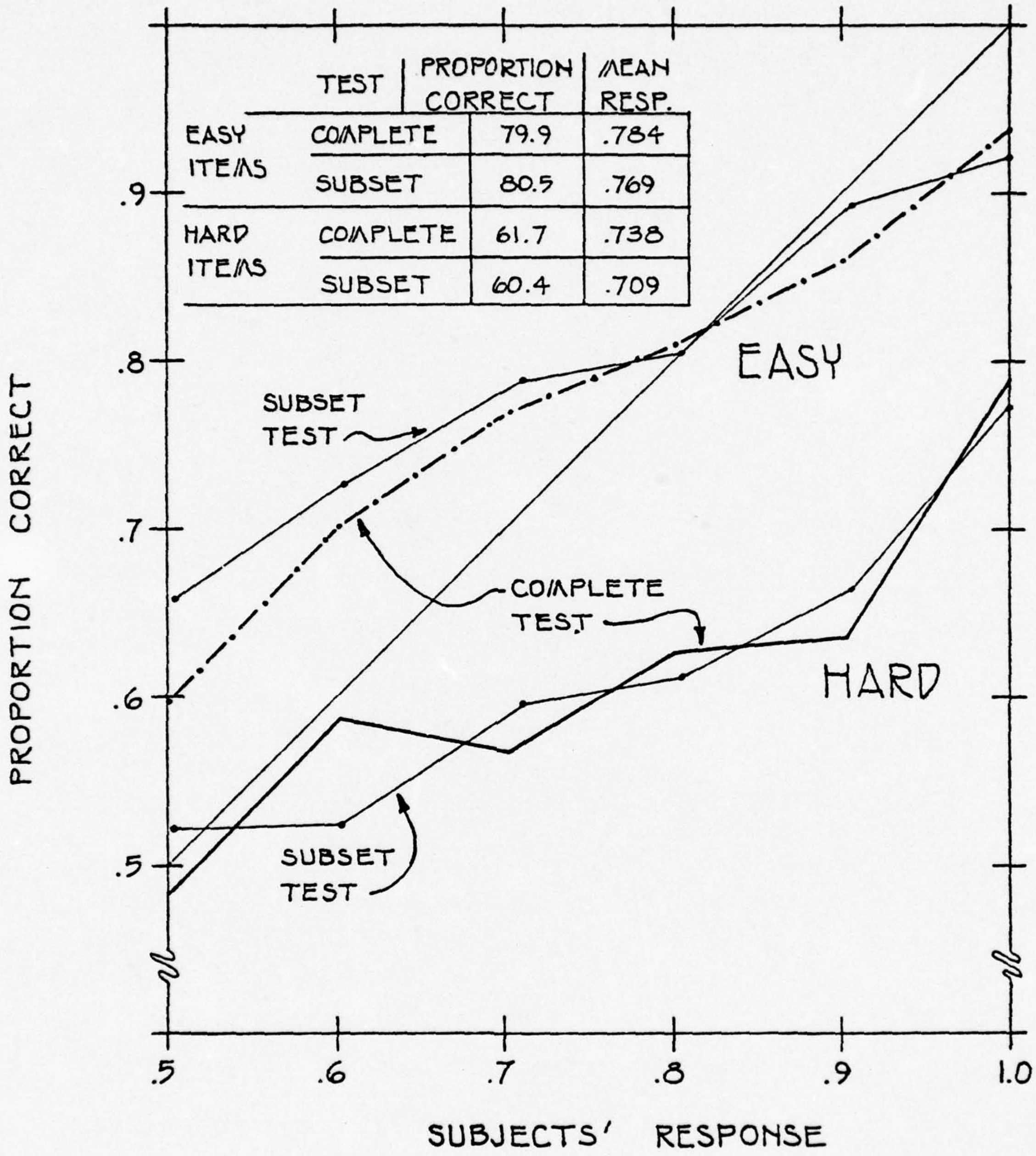
Experiment 5

Method. From the items used in Experiment 3, two tests of 50 items each were compiled. Items were selected in pairs according to the percent of subjects answering them correctly in Experiment 3. Each item in the hard test was matched with an item in the easy test that had been answered correctly by an additional 20% of subjects. The mean percent correct for the hard test was 60.4 (range, 46.2 to 77.5); for the easy test, 80.5 (range, 66.2 to 97.5).

The two tests were distributed to 93 subjects; 48 received the hard test and 45 the easy test.

Results. Figure 9 compares results from this experiment (the "complete" tests) with those from Experiment 3 using the same items (the "subset" tests). Here, too, the calibration curve depends on test difficulty, with underconfidence on the easy test and overconfidence on the hard test. The similarity between the calibration curves for the complete tests and the subset tests suggests that artifactual explanations of the results of Experiment 3 are untenable.

Eleven items of intermediate difficulty were used in both the hard and the easy tests of Experiment 5 (these were the hardest of the easy test and the easiest of the hard test). Analyses for these items revealed no



	TEST	PROPORTION CORRECT	MEAN RESP.
EASY ITEMS	COMPLETE	79.9	.784
	SUBSET	80.5	.769
HARD ITEMS	COMPLETE	61.7	.738
	SUBSET	60.4	.709

Figure 9

Complete test versus subset of a test

differences in percent correct, mean response, or calibration between the two groups. Thus, there appeared to be no context effects in responses to these items.

Expertise

Perhaps the categorization of items into "hard" and "easy" does not really capture the essence of expertise. Experts might be better calibrated not only because they know the correct answer for more of the items, but also because they have thought more about the whole topic area, and thus can more readily recognize the extent and the limitations of their knowledge. The following analysis searched for differences in calibration due to any sort of "quality of insight" that experts might have above and beyond their level of knowledge.

Method. The experts were the 50 graduate students in the Department of Psychology mentioned in the description of Experiment 4. This experiment is simply a re-analysis of that data, comparing their calibration on the 50 items pertaining to psychological knowledge with the 50 general-knowledge items.

Results. The psychology subtest and general-knowledge subtest were virtually identical in percent correct (75.7 vs. 76.0) and mean probability response (.780 vs. .778). Figure 10 shows that calibration for the two subtests was essentially the same.[2] Thus, with equal knowledge there is no evidence that expertise in a particular subject area leads to better calibration.

[2] No test of significance of the difference between two calibration curves is known. However, even the most extreme difference in Figure 10 (associated with the responses .80 to .89) has a probability of .09 of being due to chance, assuming a uniform prior.

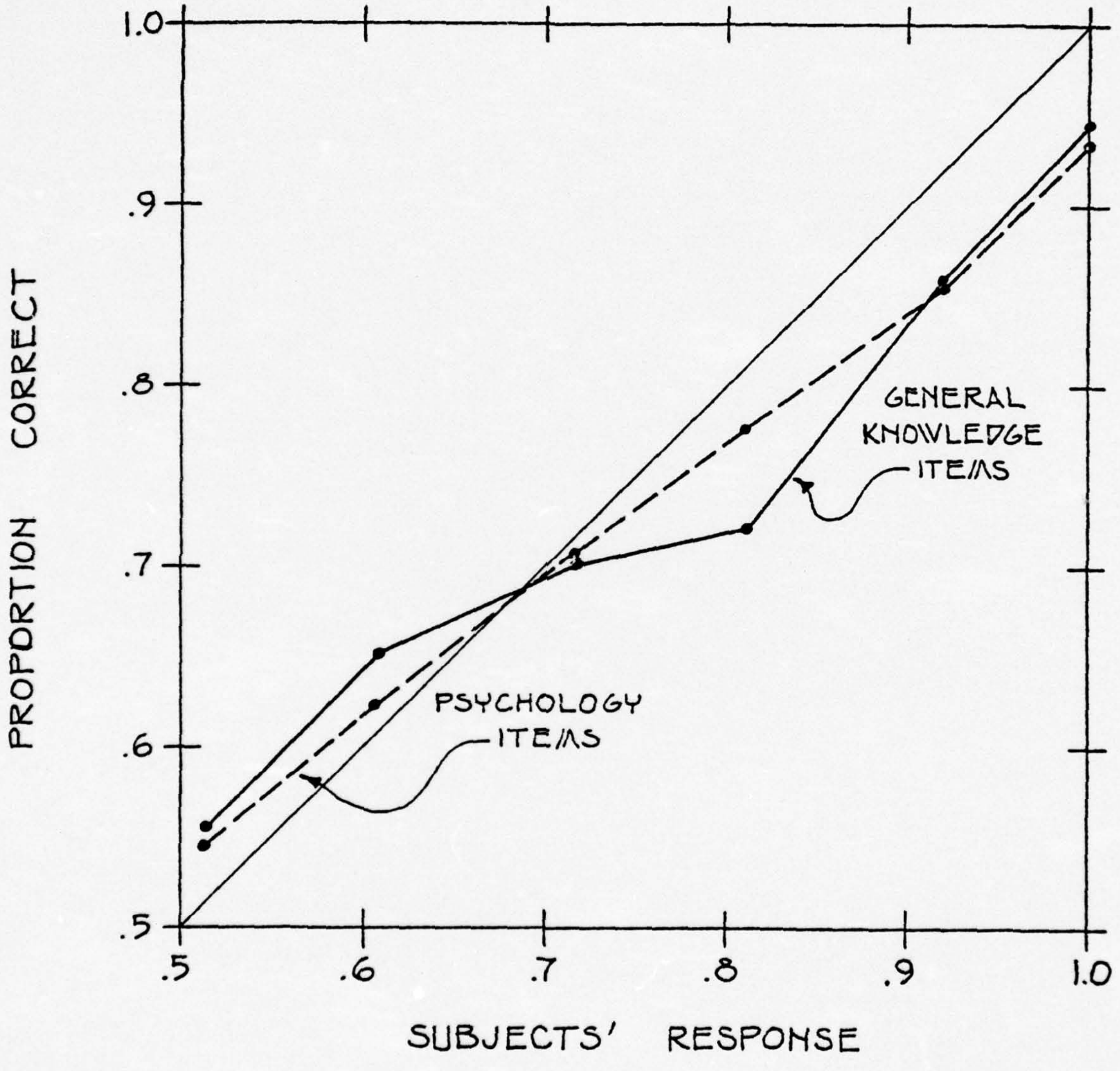


Figure 10

Effects of expertise

Intelligence

The subjects in Experiment 3 were mostly undergraduate students attending the University of Oregon. They are probably less intelligent, on the average, than the graduate student subjects of Experiment 4, who are highly selected for intelligence by the admissions procedures of the Psychology Department. We are thus able to investigate the effects of intelligence on calibration.

Method. Subtests of 73 items each, matched item by item in difficulty, were created from the Experiment 3 (regular volunteer subjects) and Experiment 4 (graduate student subjects) data.

Results. Thirty items were common to both groups. Responses to them revealed the graduate students' superior knowledge. They averaged 76.2 percent correct on these items, compared with the regular volunteers' mean of 63.9 percent correct. The graduate students had fewer correct answers on only 4 of the 30 items.

The matching process succeeded in producing subtests with a mean percent correct of 69.8 for the graduate students and 69.2 for the regular volunteers. Mean probability responses were .747 and .751, respectively.

Figure 11 shows the calibration of the two groups. It appears that the graduate students may be slightly better calibrated at the extremes. The differences, however, seem slight when compared with differences in calibration due to test difficulty.

Distribution of Responses

Figure 12 presents the proportion of subjects' probability responses that fell into each response category. These proportions are shown for

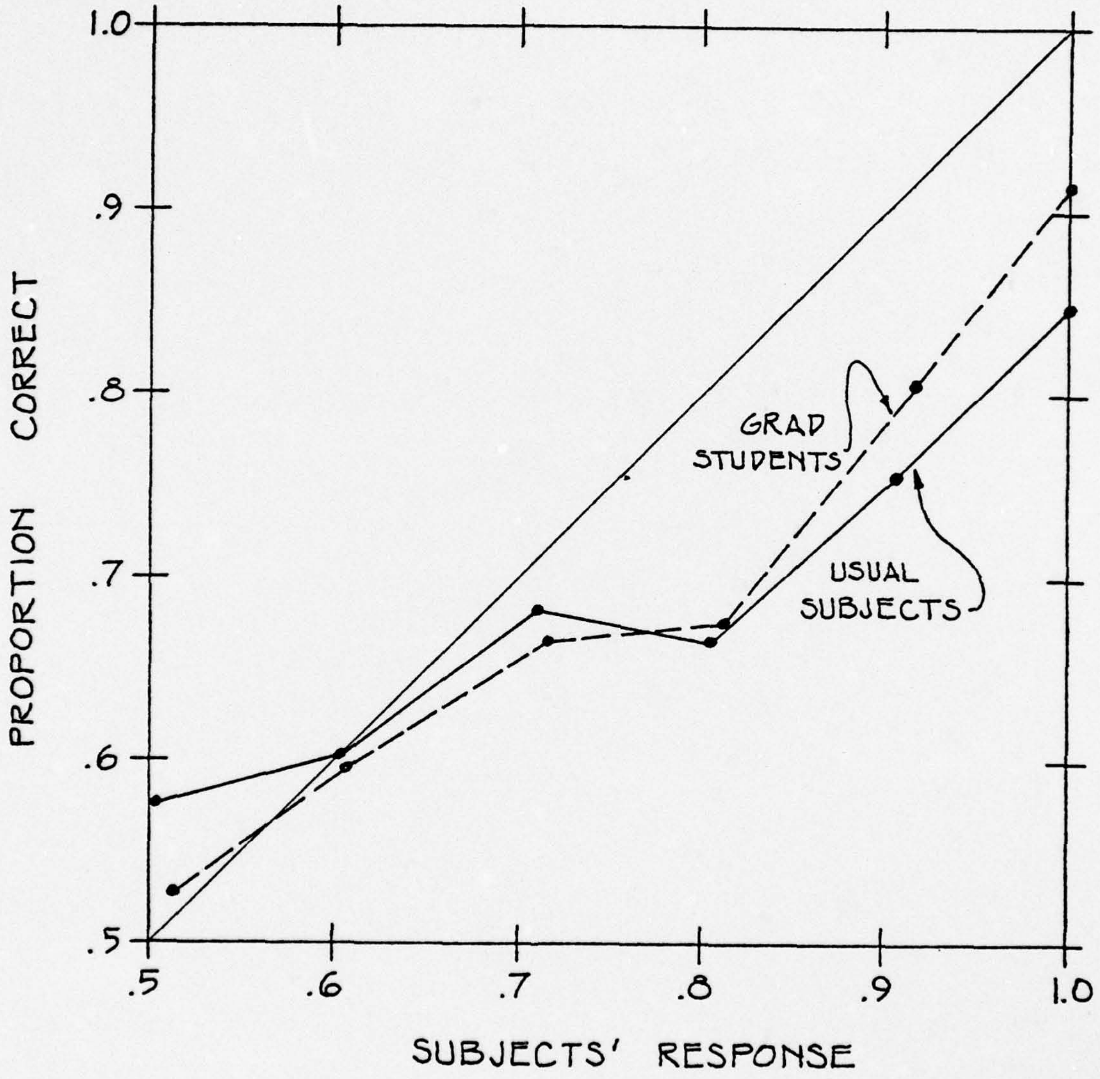


Figure 11

Effects of brains

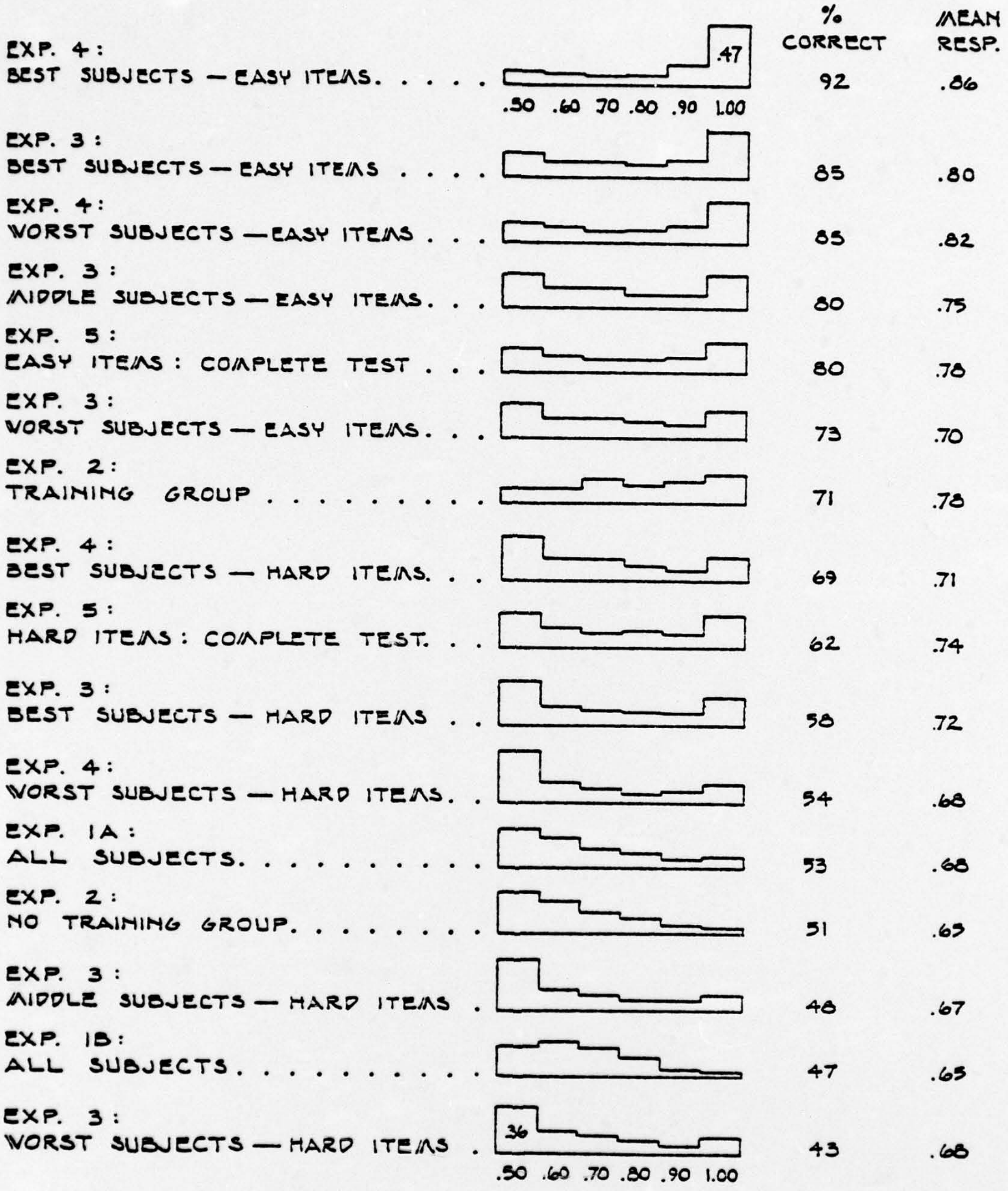


Figure 12

Distributions of subjects' responses

all groups or subgroups of all experiments, ordered by percent correct. Subjects showed a definite tendency to make more use of the high end of the response scale for the easiest tests. However, this tendency, while in the right direction, was less than it should have been. While the percent correct ranged from 43 to 92, the range of mean probability was only .65 to .86. It is this insufficient discrimination which leads to underestimation with easy tests and overestimation with hard tests.

The other striking attribute of Figure 12 is the great frequency of extreme responses (.5 and 1.0). While no response category was unused, over all experiments, subjects used the extreme categories for about half their responses. This inclination to treat the task as dichotomous (either "I know the answer"--1.0, or "I don't know the answer"--.50) appears to have been less pronounced in Experiments 1 and 2, with relatively few items all dealing with the same topic, than in Experiments 3, 4, and 5, which used many items concerning diverse topics.

The effect on calibration of the tendency to avoid using probabilities other than .5 and 1.0 was examined with the data of Experiment 3. Subjects were divided into three groups: heavy users of .5 and 1.0 (49 subjects using these two responses more than 50% of the time; mean use, 67.7%); medium users of .5 and 1.0 (33 subjects using .5 and 1.0 between 41% and 49% of the time; mean use, 46.9%); and light users (38 subjects using .5 and 1.0 40% or less of the time; mean use, 34.4%). The three groups were similar in percent of items answered correctly (65%, 64%, 62%, respectively). Their calibration curves (not shown) were highly similar, and all three groups showed the same gross overconfidence with hard items and mild underconfidence with easy items. We thus found no support for the notion that the tendency to avoid extreme probability responses, as an individual difference, affects calibration.

Discussion

At the outset, we must caution the reader about some limitations on the generalizability of these findings:

1) All subjects were naive about probabilities, and received only minimal training via experimental instructions. Even modest additions to the instructions might lead to pronounced changes in calibration.[3]

2) The items always had two alternatives, and the subjects were restricted to probabilistic responses greater than or equal to .5. Use of true-false or multi-alternative items, or elicitation of the full range of probabilities, could affect calibration.

3) Because of the large amounts of data needed for stable estimation of calibration curves, only group results are reported here. It seems reasonable that important individual differences exist in calibration, but this possibility has so far received only the most rudimentary exploration (Adams & Adams, 1961).

Nonetheless, strong effects emerged. People do show some realism and sensitivity in their probability assessments, although in general they are not well calibrated. With difficult items, assessors are overconfident; with easy items, they are underconfident.

[3] In a recent study (Slovic, Fischhoff, & Lichtenstein, 1976), we found little difference in the calibration of odds responses produced with minimal and with extensive instructions.

The strikingly different calibration curves for items of varying difficulty are a direct result of subjects' insensitivity to how much they really know. Among the items for which they believe that they have a 50% chance of knowing the correct answer, the appropriate probability may be anywhere between .45 and .85. When they estimate 1.00, the appropriate probability may be between .55 and .95 (Figure 6). The ease with which the different calibration curves were constructed from the fairly representative sets of items used in Experiments 3, 4, and 5, and the large numbers of responses in each category for even the most extreme curves, indicate that subjects' inability to make discriminations is widespread (i.e., there are not only some instances in which, for example, people should be saying .75 when they actually say .50, but many such instances).

Although subjective probabilities have a prominent role in many psychological theories, the study of *probabilities* themselves has been atheoretical in most cases (including the present study; see also Lichtenstein, Fischhoff, and Phillips, 1976). While there have been some suggestions for, or fragments of, process theories of calibration (Pitz, 1974; Slovic, 1972; Tversky & Kahneman, 1974), only Pitz (1974) predicts a decrease in overconfidence as knowledge increases.

Practical Implications. Aside from their theoretical import for the psychologist interested in how people perform judgments under conditions of uncertainty, these results have strong implications for those whose jobs involve actually making and taking responsibility for such judgments. With the development of sophisticated information processing and decision analytic techniques, operations as diverse as intelligence analysis, corporate planning, environmental impact assessment and nuclear power engineering utilize explicit probability assessments (Fischhoff, 1976). Users of these approaches should consider results like the present ones in determining how much faith to put in the results of their analyses. Similarly, psychologists who elicit subjective probability estimates in the study of behavioral phenomena might think twice before taking them at face value--or expecting too much of them.

In addition to their cautionary value, these results may also help improve the quality of probabilistic analyses. Assume that in the context of a practical problem using judgments of the type studied here, a judge reports a probability of .90. From Figure 4, we know that a better estimate of the appropriate probability is .71, and would do better treating it as such. Although such "correction after the fact" is better than taking biased judgments at face value, the revised assessments may still be inappropriate. In the present example, even though our best guess of the appropriate probability is .71, anything between .40 and .90 might be even better, depending on the difficulty of the item involved.

If we know how difficult the item is, then we can make a much more accurate correction. In practice, however, such situations will be rare. To know how difficult an item is, we must know the correct answer. But if we know the correct answer, we will not have any practical need for the judge's assessment. Such assessments are valuable only when the correct answer is not known. Short of knowing the correct answer, the only way to capitalize on the relationship between item difficulty and type of miscalibration seems to be to assume something about the difficulty of the items in the world in which our judge is functioning. The distribution of judges' responses (as shown in Figure 12) could be exploited for this purpose. Across the 16 groups or subgroups, there is a correlation of .91

between percent correct (an index of difficulty which is typically unknown in a practical setting) and mean response (which is observable when a number of assessments are made). Thus inferences about task difficulty could be made when true outcomes are unknown. With some idea of task difficulty, even so indirectly measured, more precise external recalibration of probability assessments is possible. Without it, the present data suggest that we have only a vague idea of whether to recalibrate an assessment by increasing or decreasing it.

In view of these difficulties in recalibration, it is important for future research to explore the possibility that judges can be trained to be better calibrated, thus obviating the need for correction.

REFERENCES

Adams, J. K., & Adams, P. A. Realism of confidence judgments. *Psychological Review*, 1961, *68*, 33-45.

Atomic Energy Commission. Reactor Safety Study: An assessment of accident risks in U. S. commercial power plants, WASH-1400 Draft, Washington, D. C.: The Commission, 1974.

Clarke, F. R. Confidence ratings, second-choice responses, and confusion matrices in intelligibility tests. *Journal of the Acoustical Society of America*, 1960, *32*, 35-46.

Cohen, J. *Chance, skill and luck: The psychology of guessing and gambling*. Baltimore, Md.: Penguin, 1960.

Edwards, W., & Tversky, A. *Decision making*. Baltimore, Md.: Penguin, 1967.

Feather, N. T. Subjective probability and decision under uncertainty. *Psychological Review*, 1959, *66*, 150-163.

Fischhoff, B. Cost-benefit analysis and the art of motorcycle maintenance. *Oregon Research Institute Research Monograph*, 1976, *16*, 1.

Fishbein, M. A behavior theory approach to the relations between beliefs about an object and the attidude toward the object. In M. Fishbein (Ed.), *Readings in attitude theory and measurement*. New York: John Wiley, 1967. Pp. 389-399.

Jones, E. E., & Davis, K. E. From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology, Vol. 2.* New York: Academic Press, 1965.

Kellogg, R. *Analyzing children's art*. Palo Alto, Calif.: National Press, 1970.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. Foundations of measurement, Vol. 1. New York: Academic Press, 1971.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. Calibration of probabilities: The state of the art. In H. Jungermann & G. de Zeeuw (Eds.), Proceedings of the fifth conference on subjective probability, utility and decision making. Darmstadt, Germany: 1976, in press.

Peterson, C. R., & Beach, L. R. Man as an intuitive statistician. Psychological Bulletin, 1967, 68, 29-46.

Pitz, G. F. Subjective probability distributions for imperfectly known quantities. In L. W. Gregg (Ed.), Knowledge and cognition. New York: John Wiley, 1974. Pp. 29-41.

Pollack, I., & Decker, L. R. Confidence ratings, message reception, and the receiver operating characteristic. Journal of the Acoustical Society of America, 1958, 30, 286-292.

Raiffa, H. Decision analysis. Reading, Mass.: Addison Wesley, 1968.

Slovic, P. From Shakespeare to Simon: Speculations--and some evidence--about man's ability to process information. Oregon Research Institute Research Monograph, 1972, 12, 2.

Slovic, P., Fischhoff, B., & Lichtenstein, S. The certainty illusion. Oregon Research Institute Research Bulletin, 1976, 16, 4.

Slovic, P., Kunreuther, H., & White, G. F. Decision process, rationality, and adjustment to natural hazards. In G. F. White (Ed.), Natural hazards, local, national and global. New York: Oxford University Press, 1974.

Tversky, A., & Kahneman, D. Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 1973, 5, 207-232.

Tversky, A., & Kahneman, D. Judgment under uncertainty. *Science*, 1974, *185*, 1124-1131.

Weiner, B. *Achievement motivation and attribution theory*. Morristown, N. J.: General Learning Press, 1974.

Wyer, R. S. *Cognitive organization and change: An information processing approach*. Potomac, Md.: Erlbaum, 1974.

Research Distribution List

Department of Defense

Assistant Director (Environment and Life Sciences)
Office of the Deputy Director of Defense Research and Engineering (Research and Advanced Technology)
Attention: Lt. Col. Henry L. Taylor
The Pentagon, Room 3D129
Washington, DC 20301

Office of the Assistant Secretary of Defense (Intelligence)
Attention: CDR Richard Schlaff
The Pentagon, Room 3E279
Washington, DC 20301

Director, Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

Director, Cybernetics Technology Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

Director, Program Management Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209
(two copies)

Administrator, Defense Documentation Center
Attention: DDC-TC
Cameron Station
Alexandria, VA 22314
(12 copies)

Department of the Navy

Office of the Chief of Naval Operations (OP-987)
Attention: Dr. Robert G. Smith
Washington, DC 20350

Director, Engineering Psychology Programs (Code 455)
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217
(three copies)

Assistant Chief for Technology (Code 200)
Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217

Office of Naval Research (Code 230)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Naval Analysis Programs (Code 431)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Operations Research Programs (Code 434)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research (Code 436)
Attention: Dr. Bruce McDonald
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Information Systems Program (Code 437)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research (ONR)
International Programs (Code 1021P)
800 North Quincy Street
Arlington, VA 22217

Director, ONR Branch Office
Attention: Dr. Charles Davis
536 South Clark Street
Chicago, IL 60605

Director, ONR Branch Office
Attention: Dr. J. Lester
495 Summer Street
Boston, MA 02210

Director, ONR Branch Office
Attention: Dr. E. Gloye and Mr. R. Lawson
1030 East Green Street
Pasadena, CA 91106
(two copies)

Dr. M. Bertin
Office of Naval Research
Scientific Liaison Group
American Embassy – Room A-407
APO San Francisco 96503

Director, Naval Research Laboratory
Technical Information Division (Code 2627)
Washington, DC 20375
(six copies)

Director, Naval Research Laboratory (Code 2029)
Washington, DC 20375
(six copies)

Scientific Advisor
Office of the Deputy Chief of Staff
for Research, Development and Studies
Headquarters, U.S. Marine Corps
Arlington Annex, Columbia Pike
Arlington, VA 20380

Headquarters, Naval Material Command (Code 0331)
Attention: Dr. Heber G. Moore
Washington, DC 20360

Headquarters, Naval Material Command (Code 0344)
Attention: Mr. Arnold Rubinstein
Washington, DC 20360

Naval Medical Research and Development Command (Code 44)
Naval Medical Center
Attention: CDR Paul Nelson
Bethesda, MD 20014

Head, Human Factors Division
Naval Electronics Laboratory Center
Attention: Mr. Richard Coburn
San Diego, CA 92152

Dean of Research Administration
Naval Postgraduate School
Monterey, CA 93940

Naval Personnel Research and Development Center
Management Support Department (Code 210)
San Diego, CA 92152

Naval Personnel Research and Development Center (Code 305)
Attention: Dr. Charles Gettys
San Diego, CA 92152

Dr. Fred Muckler
Manned Systems Design, Code 311
Navy Personnel Research and Development
Center
San Diego, CA 92152

Humaı Factors Department (Code N215)
Naval Training Equipment Center
Orlando, FL 32813

Training Analysis and Evaluation Group
Naval Training Equipment Center
(Code N-00T)
Attention: Dr. Alfred F. Smode
Orlando, FL 32813

Department of the Army

Technical Director, U.S. Army Institute for the Behavioral and Social Sciences
Attention: Dr. J.E. Uhlaner
1300 Wilson Boulevard
Arlington, VA 22209

Director, Individual Training and Performance Research Laboratory
U.S. Army Institute for the Behavioral and
and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

Director, Organization and Systems Research Laboratory
U.S. Army Institute for the Behavioral and
Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

Department of the Air Force

Air Force Office of Scientific Research
Life Sciences Directorate
Building 410, Bolling AFB
Washington, DC 20332

Robert G. Gough, Major, USAF
Associate Professor
Department of Economics, Geography and
Management
USAF Academy, CO 80840

Chief, Systems Effectiveness Branch
Human Engineering Division
Attention: Dr. Donald A. Topmiller
Wright-Patterson AFB, OH 45433

Aerospace Medical Division (Code RDH)
Attention: Lt. Col. John Courtright
Brooks AFB, TX 78235

Other Institutions

The Johns Hopkins University
Department of Psychology
Attention: Dr. Alphonse Chapanis
Charles and 34th Streets
Baltimore, MD 21218

Institute for Defense Analyses
Attention: Dr. Jesse Orlansky
400 Army Navy Drive
Arlington, VA 22202

Director, Social Science Research Institute
University of Southern California
Attention: Dr. Ward Edwards
Los Angeles, CA 90007

Perceptronics, Incorporated
Attention: Dr. Amos Freedy
6271 Variel Avenue
Woodland Hills, CA 91364

Director, Human Factors Wing
Defense and Civil Institute of
Environmental Medicine
P.O. Box 2000
Downsville, Toronto
Ontario, Canada

Stanford University
Attention: Dr. R.A. Howard
Stanford, CA 94305

Montgomery College
Department of Psychology
Attention: Dr. Victor Fields
Rockville, MD 20850

General Research Corporation
Attention: Mr. George Pugh
7655 Old Springhouse Road
McLean, VA 22101

Oceanautics, Incorporated
Attention: Dr. W.S. Vaughan
3308 Dodge Park Road
Landover, MD 20785

Director, Applied Psychology Unit
Medical Research Council
Attention: Dr. A.D. Baddeley
15 Chaucer Road
Cambridge, CB 2EF
England

Department of Psychology
Catholic University
Attention: Dr. Bruce M. Ross
Washington, DC 20017

Stanford Research Institute
Decision Analysis Group
Attention: Dr. Allan C. Miller III
Menlo Park, CA 94025

Human Factors Research, Incorporated
Santa Barbara Research Park
Attention: Dr. Robert R. Mackie
6780 Cortona Drive
Goleta, CA 93017

University of Washington
Department of Psychology
Attention: Dr. Lee Roy Beach
Seattle, WA 98195

Eclectech Associates, Incorporated
Post Office Box 179
Attention: Mr. Alan J. Pesch
North Stonington, CT 06359

Hebrew University
Department of Psychology
Attention: Dr. Amos Tversky
Jerusalem, Israel

Dr. T. Owen Jacobs
Post Office Box 3122
Ft. Leavenworth, KS 66027

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. ~~TITLE (and Subtitle)~~

6 The effect of knowledge on the calibration of probability assessments,

5. ~~TYPE OF REPORT & PERIOD COVERED~~

9 Technical rept.

6. ~~PERFORMING ORG. REPORT NUMBER~~

ORI ~~Report~~ No.: DDI-4

7. ~~AUTHOR(s)~~

10 Sarah Lichtenstein
Baruch Fischhoff

8. CONTRACT OR GRANT NUMBER(s)

Prime Contract No.:
N00014-76-C-0074
Subcontract No.: 75-030-0712

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Oregon Research Institute
P.O. Box 3196
Eugene, Oregon 97403

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS

Defense Advanced Research Projects Agency
1400 Wilson Blvd.
Arlington, VA 22209

12. ~~REPORT DATE~~

11 August 1976

13. ~~NUMBER OF PAGES~~

47

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

12 47p.

15. SECURITY CLASS. (of this report)

Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; ~~distribution unlimited~~

15 N00014-76-C-0074

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

ARPA Order-3052

18. SUPPLEMENTARY NOTES

Support for this research performed by Oregon Research Institute, was provided by the Advanced Research Projects Agency of the Department of Defense and was monitored under Contract N00014-76-C-0074 with the Office of Naval Research, under subcontract from Decisions and Designs, Inc.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Calibration
Probability Assessment
Knowledge
Expertise

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

One way to assess the validity of a set of subjective probability judgments is to examine their degree of calibration. The perfectly calibrated judge assigns probabilities so that of all the propositions assigned a probability of .XX of being true, XX% are in fact true. For example, half of the propositions given a .50 chance of being true should in fact be true. A series of experiments revealed that: (1) although people are moderately well calibrated, their probability judgments are prone to systematic biases. The

(cont on p 40)

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

390664

1B

cont fr p. 39)

most common bias is overconfidence; (2) people are differently calibrated when dealing with items of varying degrees of difficulty; (3) calibration is unaffected by differences in intelligence, expertise, subjects' reliance on extreme probability responses, and at least some aspects of the context in which items are presented. The implications of these results for decision makers are discussed.